

## IN THE CLAIMS

Claims 1, 6, 12, 13, 24 and 36 are amended herein. Claims 37-45 are newly presented. All pending claims are produced below.

1. (Currently Amended) A system for finding compounds in a text corpus, comprising:

a vocabulary comprising tokens extracted from a text corpus; and  
a compound finder ~~configured to~~ executable to iteratively identify compounds having a plurality of lengths within the text corpus, and rebuild at least part of the vocabulary based on the identified compounds having the plurality of lengths, each compound comprising a plurality of tokens, the compound finder comprising:

an iterator ~~executable configured~~ to select  $n$ -grams having a same length that is less than a length of  $n$ -grams selected during a previous iteration;

an  $n$ -gram counter ~~executable configured~~ to evaluate a frequency of occurrence for one or more  $n$ -grams having the same length in the text corpus, each  $n$ -gram comprising at least one token selected from the vocabulary; and

a likelihood evaluator ~~executable to configured to~~ determine a likelihood of collocation for one or more of the  $n$ -grams having the same length[.];  
add a subset of  $n$ -grams that satisfy at least one criterion evaluated responsive to the likelihood of collocation having a high likelihood as compounds to the vocabulary;  
and

~~rebuild at least part of rebuilding~~ the vocabulary  
based on the added subset of  $n$ -grams  
~~compounds.~~

2. (Cancelled)

3. (Currently Amended) A system according to Claim 1, wherein only some of the subset of  $n$ -grams ~~having a high likelihood that satisfy the at least one criterion~~ are added as compounds to the vocabulary.

4. (Original) A system according to Claim 1, wherein the likelihood of collocation as a likelihood ratio  $\lambda$  is computed in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where  $L(H_i)$  is a likelihood of observing  $H_i$  under an independence hypothesis,  $L(H_c)$  is a likelihood of observing  $H_c$  under a collocation hypothesis, and  $H$  is a pair of tokens.

5. (Original) A system according to Claim 4, wherein the  $L(H_c)$  is determined, comprising dividing the  $n$ -gram into  $n-1$  pairings of segments, calculating a likelihood of collocation for each pairing of segments, and selecting the maximum likelihood of collocation of the pairings as  $L(H_c)$ .

6. (Currently Amended) A method for finding compounds in a text corpus, comprising:

building a vocabulary comprising tokens extracted from a text corpus;

and

iteratively identifying compounds having a plurality of lengths within the text corpus and rebuilding at least part of the vocabulary based on the identified compounds having the plurality of lengths, each compound comprising a plurality of tokens, comprising:

selecting  $n$ -grams having a same length that is less than a length of  $n$ -grams selected during a previous iteration;  
 evaluating a frequency of occurrence for one or more  $n$ -grams having the same length in the text corpus, each  $n$ -gram comprising at least one token selected from the vocabulary;  
 determining a likelihood of collocation for one or more of the  $n$ -grams having the same length; ~~and~~  
 adding a subset of  $n$ -grams that satisfy at least one criterion evaluated responsive to the likelihood of collocation having a high likelihood as compounds to the vocabulary; and  
 rebuilding at least part of the vocabulary based on the added subset of  $n$ -grams compounds.

7. (Cancelled)

8. (Currently Amended) A method according to Claim 6, further comprising:

adding only some of the subset of the  $n$ -grams ~~having a high likelihood~~ that satisfy the at least one criterion as compounds to the vocabulary.

9. (Original) A method according to Claim 6, further comprising—  
 computing the likelihood of collocation as a likelihood ratio  $\lambda$  in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where  $L(H_i)$  is a likelihood of observing  $H_i$  under an independence hypothesis,  $L(H_c)$  is a likelihood of observing  $H_c$  under a collocation hypothesis, and  $H$  is a pair of tokens.

10. (Previously Presented) A method according to Claim 9, further comprising determining  $L(H_c)$ , comprising:

dividing the  $n$ -gram into  $n-1$  pairings of segments;

calculating a likelihood of collocation for each pairing of segments;

and

selecting the maximum likelihood of collocation of the pairings as

$L(H_c)$ .

11. (Original) A computer-readable storage medium holding code for performing the method according to Claim 6.

12. (Currently Amended) An apparatus for finding compounds in a text corpus, comprising:

means for building a vocabulary comprising tokens extracted from a text corpus; and

means for iteratively identifying compounds having a plurality of lengths within the text corpus and rebuilding at least part of the vocabulary based on the identified compounds having the plurality of lengths, each compound comprising a plurality of tokens, comprising:

means for selecting  $n$ -grams having a same length that is less than a length of  $n$ -grams selected during a previous iteration;

means for evaluating a frequency of occurrence for one or more  $n$ -grams having the same length in the text corpus, each  $n$ -gram comprising at least one token selected from the vocabulary;

means for determining a likelihood of collocation for one or more of the  $n$ -grams having the same length; and

means for adding a subset of  $n$ -grams that satisfy at least one criterion evaluated responsive to the likelihood

~~of collocation having a highest likelihood as~~  
~~compounds~~ to the vocabulary; and  
means for rebuilding at least part of the vocabulary based  
on the added subset of  $n$ -grams ~~compounds~~.

13. (Currently Amended) A system for identifying compounds through iterative analysis of measure of association, comprising:
- an iterator ~~executable to~~ initially ~~specify~~ specifying a limit on a number of tokens per compound for an iteration and decreasing the limit for a subsequent iteration; and
  - a compound finder ~~configured executable to~~ iteratively identify ~~evaluate~~ compounds having a plurality of lengths within a text corpus and rebuild at least part of a vocabulary for the text corpus based on the identified compounds having the plurality of lengths, comprising:
    - an  $n$ -gram counter ~~executable~~ configured to determine; ~~determine~~ a number of occurrences of one or more  $n$ -grams within the text corpus, each  $n$ -gram comprising a number of tokens up to the limit for the iteration, which are at least in part provided in ~~the~~ a vocabulary for the text corpus;
    - a likelihood evaluator ~~executable~~ configured to identify; ~~identify~~ identifying at least one  $n$ -gram comprising a number of tokens equal to the limit for the iteration based on the number of occurrences; ~~and~~
    - ~~determine~~ determining a measure of association between the tokens in the identified  $n$ -gram~~[[.]]~~ ;
    - add ~~adding~~ each identified  $n$ -gram with a sufficient measure of association to the vocabulary as a compound token; and

rebuild at least part of rebuilding the vocabulary based  
on the added compound tokens.

14. (Previously Presented) A system according to Claim 13, further comprising:  
a stored upper limit on a number of identified  $n$ -grams; and  
a limiter identifying a number of  $n$ -grams up to the stored upper limit  
based on the number of occurrences.

15. (Cancelled)

16. (Original) A system according to Claim 13, wherein the measure of association between the tokens in the identified  $n$ -gram comprises a likelihood ratio  $\lambda$ .

17. (Original) A system according to Claim 16, wherein the likelihood ratio  $\lambda$  is calculated in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where  $L(H_i)$  is a likelihood of observing  $H_i$  under an independence hypothesis,  $L(H_c)$  is a likelihood of observing  $H_c$  under a collocation hypothesis, and  $H$  is a pair of tokens.

18. (Original) A system according to Claim 17, wherein, for each pair of tokens,  $t_1, t_2$ , in the identified  $n$ -gram, the independence hypothesis comprises  $P(t_2 | t_1) = P(t_2 | \bar{t}_1)$  and the collocation hypothesis comprises  $P(t_2 | t_1) > P(t_2 | \bar{t}_1)$ .

19. (Original) A system according to Claim 17, wherein the  $L(H_i)$  is computed for each pair of tokens,  $t_1, t_2$ , in the identified  $n$ -gram in accordance with the formula:

$$\arg \max_{L(H_i)} \frac{L(t_1, t_2 \text{ form compound})}{L(n - \text{gram does not form compound})}.$$

20. (Original) A system according to Claim 13, further comprising:  
an initial vocabulary comprising a plurality of tokens extracted from  
the text corpus.
21. (Original) A system according to Claim 20, further comprising:  
a parser parsing the tokens from the text corpus.
22. (Original) A system according to Claim 13, further comprising:  
a filter determining the number of occurrences of one or more  $n$ -grams  
within the text corpus for only unique  $n$ -grams.
23. (Original) A system according to Claim 13, wherein each text  
corpus comprises a plurality of documents comprising one of a Web page, a news  
message and text.
24. (Currently Amended) A method for identifying compounds  
through iterative analysis of measure of association, comprising:  
iteratively specifying a limit on a number of tokens per compound for  
an iteration and decreasing the limit for a subsequent iteration;  
and  
iteratively ~~identifying, evaluating~~ compounds having a plurality of  
lengths within a text corpus and rebuilding at least part of a  
vocabulary comprised of tokens from a text corpus based on  
the identified compounds having the plurality of lengths.  
comprising:  
determining a number of occurrences of one or more  $n$ -  
grams within the text corpus, each  $n$ -gram  
comprising up to a number of tokens up to the limit  
for the iteration, which are at least in part provided  
in a the vocabulary ~~for the text corpus~~;

identifying at least one  $n$ -gram comprising a number of tokens equal to the limit for the iteration based on the number of occurrences and determining a measure of association between the tokens in the identified  $n$ -gram;

adding each identified  $n$ -gram with a sufficient that satisfies at least one criterion evaluated responsive to the measure of association to the vocabulary as a compound token; and;

rebuilding at least part of the vocabulary based on the added compound tokens.

25. (Original) A method according to Claim 24, further comprising: providing an upper limit on a number of identified  $n$ -grams; and identifying a number of  $n$ -grams up to the upper limit based on the number of occurrences.

26. (Cancelled)

27. (Original) A method according to Claim 24, wherein the measure of association between the tokens in the identified  $n$ -gram comprises a likelihood ratio  $\lambda$ .

28. (Previously Presented) A method according to Claim 27, further comprising calculating the likelihood ratio  $\lambda$  in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where  $L(H_i)$  is a likelihood of observing  $H_i$  under an independence hypothesis,  $L(H_c)$  is a likelihood of observing  $H_c$  under a collocation hypothesis, and  $H$  is a pair of tokens.



29. (Original) A method according to Claim 28, wherein, for each pair of tokens,  $t_1, t_2$ , in the identified  $n$ -gram, the independence hypothesis comprises  $P(t_2 | t_1) = P(t_2 | \bar{t}_1)$  and the collocation hypothesis comprises  $P(t_2 | t_1) > P(t_2 | \bar{t}_1)$ .

30. (Original) A method according to Claim 28, further comprising:  
computing the  $L(H_i)$  for each pair of tokens,  $t_1, t_2$ , in the identified  $n$ -gram in accordance with the formula:

$$\arg \max_{L(H_i)} \frac{L(t_1, t_2, \text{form compound})}{L(n - \text{gram does not form compound})}.$$

31. (Original) A method according to Claim 24, further comprising:  
constructing an initial vocabulary comprising a plurality of tokens  
extracted from the text corpus.

32. (Original) A method according to Claim 31, further comprising:  
parsing the tokens from the text corpus.

33. (Original) A method according to Claim 24, further comprising:  
determining the number of occurrences of one or more  $n$ -grams within  
the text corpus for only unique  $n$ -grams.

34. (Original) A method according to Claim 24, wherein each text  
corpus comprises a plurality of documents comprising one of a Web page, a news  
message and text.

35. (Original) A computer-readable storage medium holding code for  
performing the method according to Claim 24.

36. (Currently Amended) An apparatus for identifying compounds  
through iterative analysis of measure of association, comprising:  
means for specifying a limit on a number of tokens per compound for  
an iteration and decreasing the limit for a subsequent iteration;  
and

means for iteratively identifying evaluating compounds having a plurality of lengths within a text corpus and rebuilding at least part of a vocabulary comprised of tokens from a text corpus based on the identified compounds having the plurality of lengths, comprising:

means for determining a number of occurrences of one or more  $n$ -grams within the text corpus, each  $n$ -gram comprising up to a number of tokens up to the limit for the iteration, which are at least in part provided in a vocabulary for the text corpus;

means for identifying at least one  $n$ -gram comprising a number of tokens equal to the limit for the iteration based on the number of occurrences and means for determining a measure of association between the tokens in the identified  $n$ -gram; and

means for adding each identified  $n$ -gram with a sufficient that satisfies at least one criterion evaluated responsive to the measure of association to the vocabulary as a compound token and means for rebuilding at least part of the vocabulary based on the added compound tokens.

37. (New) The system of claim 1, wherein the added subset of  $n$ -grams satisfy a criterion of having a highest likelihood of collocation.

38. (New) The system of claim 37, wherein a number of  $n$ -grams in the added subset of  $n$ -grams is equal to a defined number which specifies a maximum number of  $n$ -grams having a highest likelihood of collocation to be added.

39. (New) The system of claim 1, wherein the likelihood of collocation for each  $n$ -gram of the added subset of  $n$ -grams satisfy a criterion of exceeding a threshold likelihood of collocation.

40. (New) The method of claim 6, wherein the added subset of  $n$ -grams satisfy a criterion of having a highest likelihood of collocation.

41. (New) The method of claim 40, wherein a number of  $n$ -grams in the added subset of  $n$ -grams is equal to a defined number which specifies a maximum number of  $n$ -grams having a highest likelihood of collocation to be added.

42. (New) The method of claim 6, wherein the likelihood of collocation for each  $n$ -gram of the added subset of  $n$ -grams satisfy a criterion of exceeding a threshold likelihood of collocation.

43. (New) The apparatus of claim 12, wherein the added subset of  $n$ -grams satisfy a criterion of having a highest likelihood of collocation.

44. (New) The apparatus of claim 43, wherein a number of  $n$ -grams in the added subset of  $n$ -grams is equal to a defined number which specifies a maximum number of  $n$ -grams having a highest likelihood of collocation to be added.

45. (New) The apparatus of claim 12, wherein the likelihood of collocation for each  $n$ -gram of the added subset of  $n$ -grams satisfies a criterion of exceeding a threshold likelihood of collocation.

46. (New) The system of claim 13, wherein the added subset of  $n$ -grams satisfy a criterion of having a highest measure of association.

47. (New) The system of claim 46, wherein a number of  $n$ -grams in the added subset of  $n$ -grams is equal to a defined number which specifies a maximum number of  $n$ -grams having a highest measure of association to be added.

48. (New) The system of claim 13, wherein the likelihood of collocation for each  $n$ -gram of the added subset of  $n$ -grams satisfies a criterion of exceeding a threshold measure of association.

49. (New) The method of claim 24, wherein the added subset of  $n$ -grams satisfy a criterion of having a highest measure of association.

50. (New) The method of claim 49, wherein a number of  $n$ -grams in the added subset of  $n$ -grams is equal to a defined number which specifies a maximum number of  $n$ -grams having a highest measure of association to be added.

51. (New) The method of claim 24, wherein the likelihood of collocation for each  $n$ -gram of the added subset of  $n$ -grams satisfies a criterion of exceeding a threshold measure of association.

52. (New) The apparatus of claim 36, wherein the added subset of  $n$ -grams satisfy a criterion of having a highest measure of association.

53. (New) The apparatus of claim 52, wherein a number of  $n$ -grams in the added subset of  $n$ -grams is equal to a defined number which specifies a maximum number of  $n$ -grams having a highest measure of association to be added.

54. (New) The apparatus of claim 36, wherein the likelihood of collocation for each  $n$ -gram of the added subset of  $n$ -grams satisfies a criterion of exceeding a threshold measure of association.